CalibNet: 双分支跨模态校准的 RGB-D 显著实例分割

裴佳伦，*Member, IEEE*，蒋韬，唐赫，刘念，*Member, IEEE*，金 Yue 明，*Member, IEEE*，范登平，*Senior Member, IEEE*，Pheng-Ann Heng，*Senior Member, IEEE*

摘要—本文提出了一个全新的 RGB-D 显著实例分割模型 *CalibNet*，该模型使用双分支跨模态特征校准架构，能够同时在卷积核分支和掩码分支校准 RGB 和深度特征，生成实例感知的卷积核和掩码特征。*CalibNet* 包含了三个简洁的模块。其中，动态交互的卷积核模块（DIK）和共享权重的融合模块（WSF）一起能够生成有效的实例感知卷积核并整合跨模态特征。本文还在 DIK 和 WSF 之前嵌入了深度相似性评估模块（DSA），来提升深度特征的质量。除此之外，本文进一步构建了新的数据集 DSIS，其中包括了共计 1940 组图像，并带有细致标注的实例级标签。在三个具有挑战性的评价基准上的大量实验表明 *CalibNet* 取得了优异的性能，例如在 COME15K-E 测试集上输入大小为 320×480 时取得了 58.0%AP，大幅领先于现有的显著实例分割框架。DSIS 数据集与 *CalibNet* 代码公开于https://github.com/PJLallen/CalibNet。

关键词—RGB-D 显著实例分割，跨模态融合，实例级分割，Transformer.

I. 引言

在 RGB-D 显著目标检测（SOD）任务中，深度图被广泛地用作为提供直观空间结构线索的额外信息源 [1], [2]。为了有效地利用深度信息，大部分的 RGB-D SOD 方法 [3]–[5] 在主干网络或解码器中采用跨模态特征融合模块以及深度质量评估策略，但将深度模态加入实例级显著性检测的方法却相对有限。与 SOD 相比，显著实例分割（SIS）能够确定每个显著实例的位置和数量，这对于自动驾驶 [6]、视频监控 [7] 和场景理解 [8] 等实际应用都具有很大的价值。因此，本文首次尝试引入深度线索来提升实例级显著性分割任务中的 RGB 特征表示，构建了 RGB-D SIS 任务（如图 1所示）。

图 1. RGB-D 显著实例分割任务与所提出 *CalibNet* 的预测结果示意图。本文的方法将 RGB-D 显著性检测拓展至实例级。

目前有两种典型的 RGB 实例级分割框架：基于 RPN 的框架 [10], [11]（*i.e.*，两阶段策略）以及基于类别 & 掩码的框架 [12], [13]（*i.e.*，单阶段策略）。两阶段 RGB-D 框架 (图 2 (a)) 需要额外的区域选取以及细化过程，这会增加计算复杂度和时间开销。因此，本文拟采用单阶段架构来有效地融合多尺度 RGB 和深度特征，并提取跨模态特征共享信息，生成精细的实例嵌入。然而，单阶段 RGB-D 方案可能难以应对一些深度和 RGB 信息不一致或存在噪声的情况，从而产生不准确的实例分割结果。

基于上述分析，本文提出了一个双分支跨模态校准网络 ***CalibNet***，其能够通过单阶段的方式识别显著实例。如图 2 (b) 所示，与两阶段 RGB-D 实例级分割模型的候选区域融合相比，*CalibNet* 采用了双向融合策略在卷积核分支以及掩码分支整合深度特征。双分支跨模态校准的关键在于有效利用深度线索来增强两个分

Tao Jiang, He Tang are with the School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China (e-mail: jtao99@hust.edu.cn; hetang@hust.edu.cn).

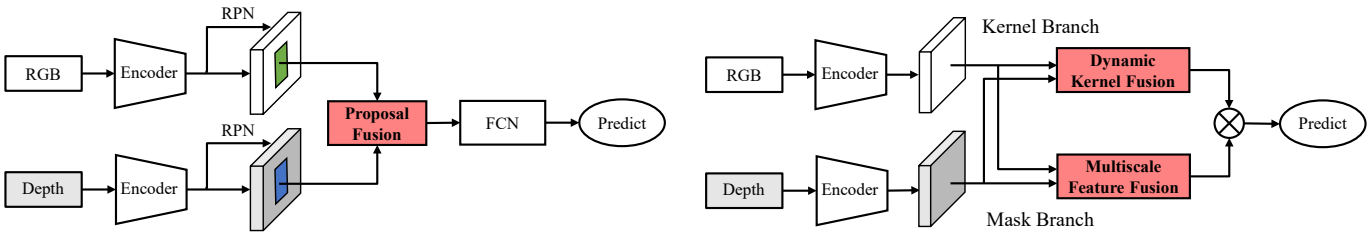
Jialun Pei, Pheng-Ann Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, HKSAR, China (e-mail: jialunpei@cuhk.edu.hk; pheng@cse.cuhk.edu.hk).

Nian Liu is with the Computer Vision Department, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: liunian228@gmail.com).

Yueming Jin is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore (e-mail: ymjin@nus.edu.sg).

Deng-Ping Fan is with Nankai International Advanced Research Institute (Shenzhen Futian),Nankai University, Shenzhen, China and he is also with College of Computer Science, Nankai University, Tianjin, China (e-mail: dengpfan@gmail.com).

Corresponding author: Deng-Ping Fan.



(a) 候选区域融合 (b) 双分支融合（本文的）

图 2. 两种 RGB-D 实例级分割模型的融合架构比较。(a) 两阶段方式的候选区域融合 [9]；(b) 本文采用的单阶段的双分支融合。

支中的多尺度 RGB 特征，从而确保实例感知卷积核与生成的掩码特征之间具较强的一致性。为此，本文在卷积核分支中引入了动态交互卷积核（**DIK**）模块，利用 RGB 和深度特征生成信息丰富的结构注意力图，并结合全局特征生成动态卷积核。在掩码分支中，主要的挑战在于抑制深度特征中的低质量结构信息，并将有价值的空间线索与 RGB 特征相融合。为了解决这一问题，本文提出了一种共享权重的融合（**WSF**）模块，它能够有效地利用跨模态共同信息，整合多尺度共享特征来增强掩码特征表示。此外，模型中还嵌入了深度相似性评估（**DSA**）模块，以识别有价值的补充信息，减少低质量的深度信息对分割结果造成的影响。本文的方法旨在提高掩码和卷积核分支中使用的深度信息的准确性和可靠性，从而获得更为精确的分割结果。此外，本文不依赖于非极大值抑制（NMS）等后处理操作，而是使用二分匹配 [14] 机制来进行标签分配并去除冗余的预测，能够在取得更好性能的同时减轻框架整体的运算量。

此外，目前仅有两个数据集可以用于 RGB-D SIS 任务，包括 COME15K [15] 和 SIP [16]。其中，COME15K 数据集的训练样本最多，但其标签质量欠佳；SIP 有着精确的标签，但其只包含 929 张图像，且仅涵盖了“人”这一类显著目标。为了推动 RGB-D SIS 任务的发展并更好地评估模型的泛化性能，有必要构建一个规模适中、高质量的标注且包含多个类别显著实例的测试集。因此，本文构建了一个名为 **DSIS** 的数据集，其中包含了 1,940 张图像，并带有细致的实例级标签。为了评估现有先进方法的性能，本文进一步在三个 RGB-D SIS 数据集上评估了 *CalibNet*，包括 COME15K、SIP 以及本文提出的 DSIS，并进行了 RGB-D SIS 领域第一个系统性的调研。

本文的主要贡献有以下三点：

- 据我们所知，***CalibNet*** 是第一个 RGB-D SIS 框架。该单阶段架构充分挖掘了深度信息以用于卷积核和掩码分支与 RGB 特征交互并优化 RGB 特征，三个 RGB-D SIS 数据集上的结果表明 *CalibNet* 具有很好的性能。
- 本文的框架中包含了动态交互卷积核（DIK）模块用于突出重要的空间位置并动态生成实例级卷积核。此外，本文引入了共享权重融合（WSF）模块来提升多尺度深度和 RGB 特征的融合效率。为了校准深度特征，本文还提供了深度相似性评估（DSA）模块来挖掘有价值的深度信息。
- 本文构建了一个新的数据集 **DSIS** 来更好地评估 RGB-D SIS 方法的泛化性能。基于 DSIS 和现有的两个数据集，本文进行了第一个大规模 RGB-D SIS 调研，评估了共计 19 种不同基线方法的性能以推动 RGB-D SIS 研究的发展。

II. 相关工作

A. RGB-D 显著目标检测

深度模态包含了丰富的几何信息，在显著性检测任务尤其是更具挑战性的场景中起到了重要作用。得益于卷积神经网络出色的特征表示能力，现有的 RGB-D SOD 方法 [17]–[21] 利用多级 RGB 和深度特征以取得出色的分割效果。这些模型主要基于全卷积神经网络（FCNs）[22] 来设计跨模态融合，其可以分为三类：早期融合 [17], [23], [24]、后期融合 [25]–[27] 以及中间层多级融合 [28]–[35]。早期融合方式首先将 RGB 图像和深度图按通道进行拼接，然后共同送入一个统一的编码器进行处理。与之相对的，后期融合策略先分别提取各个模态输入的特征，再在最后一层进行融合。为了加强 RGB 和深度特征之间的关联性，最近的 RGB-D SOD 方法大多采用中间层多级融合方式，逐步在不同尺度和层级对两个模态特征进行交互。例如，Fu 等人 [36] 采用带有联合学习模块和合作融合模块的孪生结构来提取多层级特征，挖掘跨模态互补性。Wu 等人 [37] 设计

了一个高效的分割网络来改进特征表示，并在不产生高计算开销的情况下融合 RGB 和深度特征。后来，Liu 等人 [2] 设计出一种纯 Transformer 模型，以序列-序列的方式提取全局特征，并通过交叉注意力机制融合 RGB 和深度特征。然而，现有工作并没有利用深度数据来增强实例级 SOD 模型的特征表示。因此，我们尝试针对 RGB-D SIS 任务设计一种新的跨模态融合框架。

B. 显著实例分割

SIS 的目的是识别显著区域内的所有实例，并确定显著实例数量以及其像素级掩码。与通用实例分割（GIS）任务 [10], [12], [38] 不同，SIS 是与类别无关的，因为识别显著实例是基于显著区域进行的，而不是特定的类别标签。Li 等人 [39] 首次提出了 SIS 任务以及第一个 SIS 数据集（ILSO），并为此设计了一个多阶段模型。此后，又有一些 SIS 模型陆续被提出，大多数基于 Mask R-CNN 结构 [10]。S4Net [11] 第一次采用了单阶段端到端的架构来分割显著实例。在这个基础上，RDPNet [40] 采用了正则化密集金字塔网络来进一步改进特征金字塔并抑制无效信息区域。相似的，SCG [41] 引入了显著性分支和轮廓分支来指导 Mask R-CNN 模型更好地识别显著实例。与全监督学习不同，Tian 等人 [42] 采用感数以及类别标签作为弱监督手段，提出了新的弱监督框架。最近，OQTR [43] 提出了一个基于 Transformer 的 SIS 模型，使用跨模态特征融合模块来获取全局特征，并引入了显著性引导查询来提升模型收敛速度。尽管上述方法推动了 SIS 领域快速发展，但目前仍局限于 RGB 图像的单模态输入。

C. 基于 *RGB-D* 的实例分割

尽管基于 RGB 图像的实例分割方法研究取得了较大进展，但基于 RGB-D 数据的实例分割呈现较慢的发展趋势。早期，Gupta 等人 [44] 和 Uhrig 等人 [45] 分别使用多阶段和全卷积网络来分割目标实例。近期，Xu 等人 [9] 和 Xiang 等人 [46] 又分别引入了 Mask R-CNN 架构和合成数据来进一步提升模型性能。其中，Xu 等人 [9] 提出了一种基于 Mask R-CNN 的端到端户外场景 RGB-D 实例分割框架，其在编码器中分别提取 RGB 和深度特征，然后采用残差连接进行特征融合。相似的，Xiang 等人 [46] 直接使用合成数据来学习 RGB-D 特征嵌入，并采用聚类算法来分割未见过的目标。为了推动 RGB-D 实例级分割的发展，本文提出了一个全新的 RGB-D SIS 架构，基于 CNN-Transformer 结构在卷积核和掩码分支中整合深度特征。

III. DSIS 数据集

为了创建一个鲁棒的 RGB-D SIS 基准并全面地评估特定方法的泛化性能，本文构建了一个全新的 RGB-D SIS 测试集，名为 DSIS。

A. 数据集统计

为了确保实例级标签和显著区域之间的高度一致性，本文从现有的 RGB-D 数据集中仔细收集了约 2,500 组带有区域级标注和深度图的图像对，其中包括 NJUD [47]、DUT-RGBD [35]、STEREE [48] 和 NLPR [49] 数据集。随后，五位标注者基于现有的区域级标签手工标注了像素级显著实例标签。在那之后，本文采用多数投票的方式 [50] 来保留多数结果的标签，同时舍弃了模样本。最后，本文共计收集了 1,940 组图像，并带有三种标注形式，包括边界框、显著区域和显著实例标注。图 4的左侧展示了数据来源的分布。为了提高 RGB-D SIS 任务评估的多样性，所收集到的样本都被用作测试集。此外，本文还统计了每张图像中显著实例的数量分布，如图 4的右侧所示。

在现有 SOD 真值的基础上，本文进行了细粒度的实例级标注以对 RGB-D SIS 模型进行更全面的评估，图 3展示了一些典型的实例标签。可以看出，DSIS 数据集涵盖了多种场景和外观多样的实例类型，包括复杂背景、光照变化、低对比度、边缘复杂等情况。在各种场景中，DSIS 都对显著实例进行了精细标注，如低对比度的飞机、光照变化下的动物、极细的弓箭等。多样的显著实例能够对 RGB-D SIS 模型进行更全面的评估。与 COME15K 数据集相比，本文所提出的 DSIS 测试集更好地保证了区域级标签与实例级标签的一致性，而 COME15K 标注中的一些显著实例区域并不在显著区域之中。与 SIP 数据集相比，DSIS 的数据集包含了更多样的场景以及更多类别的显著实例。贡献的 DSIS 数据集可以进一步促进 RGB-D SIS 领域的研究和发展。

B. 数据集分析与比较

1) 中心偏差： 实例中心的分布是评价数据集质量的关键指标，它可以揭示数据集是否存在中心偏差，即显著实例分布是否偏向场景中心。实例中心分布得越分散，表示数据集具有更强的实例多样性。图 6表

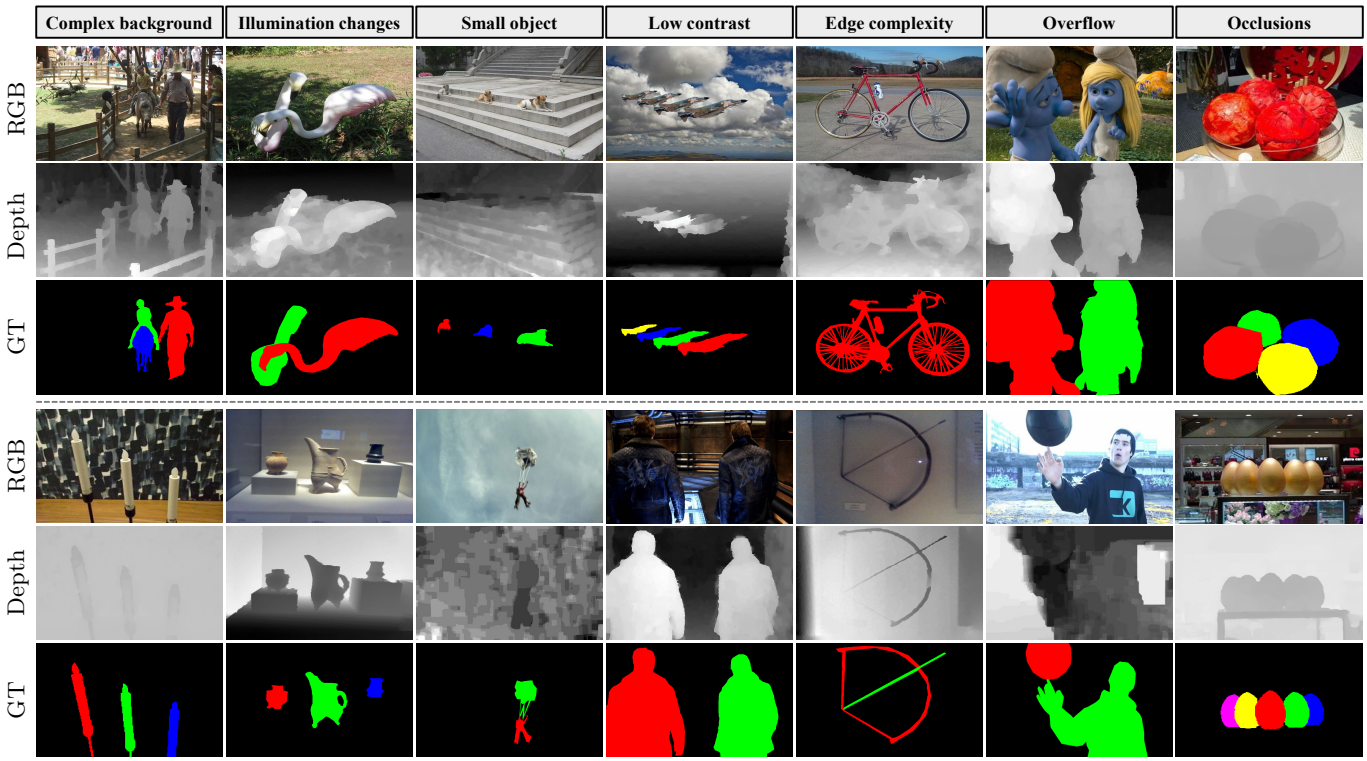


图 3. 所提出 DSIS 数据集的多样化标签示例。

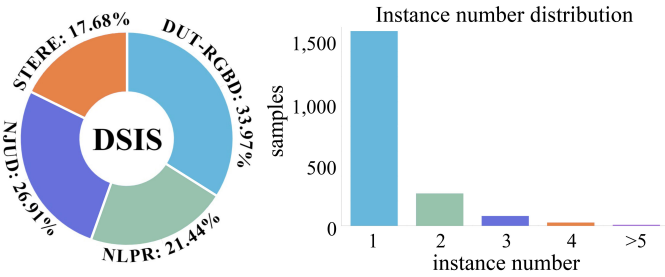


图 4. DSIS 数据集的分布。左侧：从 RGB-D SOD 数据集收集的图像来源分布。右侧：每个样本中实例数量的分布。

明所有 RGB-D SIS 数据集的实例中心都一定程度上集中于图像中心位置。具体来说，在 COME15K-E 和 COME15K-H 测试集 [15] 中，实例中心位置整体呈现中心十字分布；而 SIP 数据集 [16] 的实例中心则更多地分布在图像的中下部。相比较而言，DSIS 数据集的实例中心则在各个方向上扩散得更加均匀。

2) 实例尺寸分布： 实例尺寸分布也是评估数据集质量的重要指标，旨在反映图像中实例的变化和多样性，通常用实例像素占图像像素的比例来衡量这一指标 [51]。如图 5 (a) 所示，DSIS 和 COME15K 数据集中涵盖的实例尺寸范围更广，而 SIP 的实例尺寸变化较小。值得注意的是，DSIS 包含的中等尺寸的实例占比更高，而 COME15K 则包含较多小尺寸实例。

3) 区域-实例标签一致性： 众所周知，SIS 可以看作是在显著区域内进一步区分实例。为了确保显著实例标签的准确性，首先需要保证实例级标签与区域级标签的一致性。为了验证两种级别标签的一致性，本文将显著实例标签二值化，并计算与区域级标签的平均交并比 (IoU)。图 5 (b) 展示了区域级与实例级标签的匹配曲线，结果表明 SIP 和 DSIS 数据集中区域级和实例级标签是完全对齐的，而 COME15K 的两个测试集中平均 IoU 值都有不同程度的下降。这说明 COME15K 的显著实例标签并非完全显著的，这难以保证实例标注的质量。图 7直观地展示了 COME15K 中一些区域-实例级标签不一致的典型示例。

4) 深度-显著区域一致性： 在 RGB-D 实例级显著性检测中，深度图提供的结构信息与显著区域之间的相关性可作为衡量 RGB-D 数据集难度的一项重要指标。为此，本文对深度图进行了二值化操作，并将其与显著区域标签进行直接比较。具体来说，本文首先采用 OTSU 算法 [52] 将深度图二值化，然后再将前景和背景区域反转，以保持显著目标和空间较近目标的索引一致。图 5 (c) 展示了所有数据集中二值化后的深度图与显著区域标签的匹配曲线。可以看到，所有数据集中两

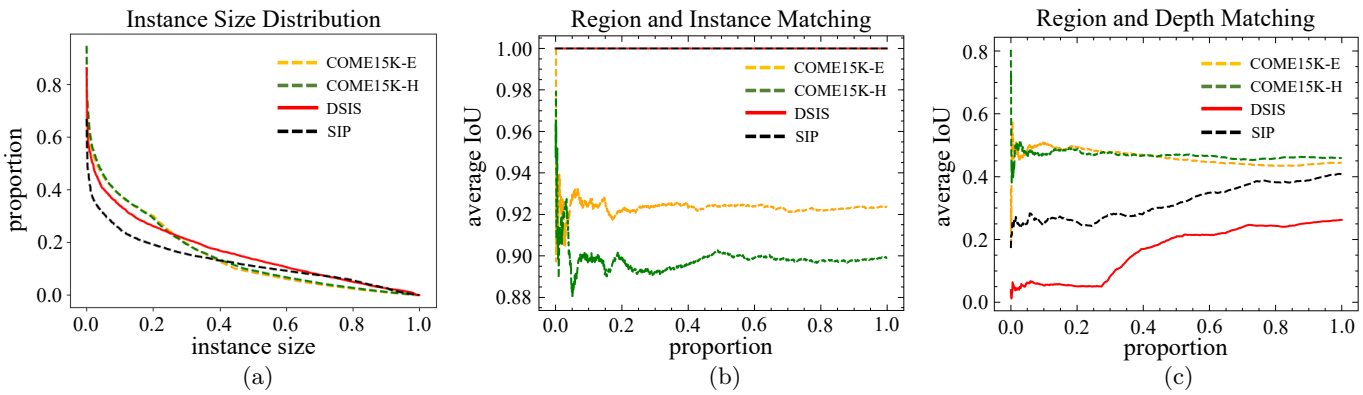


图 5. 所提出 DSIS 数据集与现有 RGB-D SIS 数据集的对比。(a) 所有测试集中实例大小的分布；(b) 显著区域真值与二值化实例真值的一致性比较；(c) 显著区域真值与二值化深度图的一致性比较。

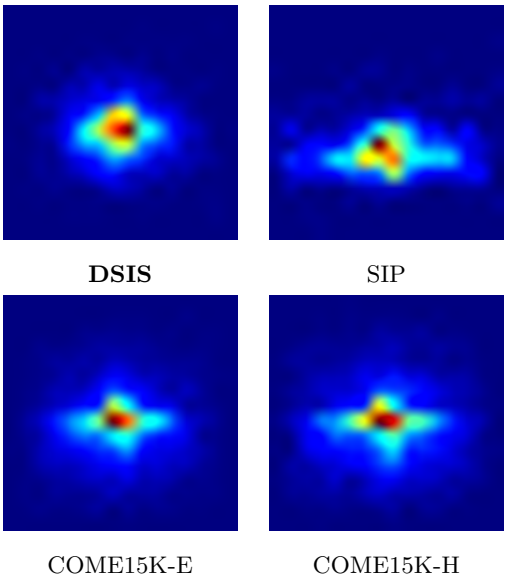


图 6. RGB-D SIS 数据集的实例中心偏差。

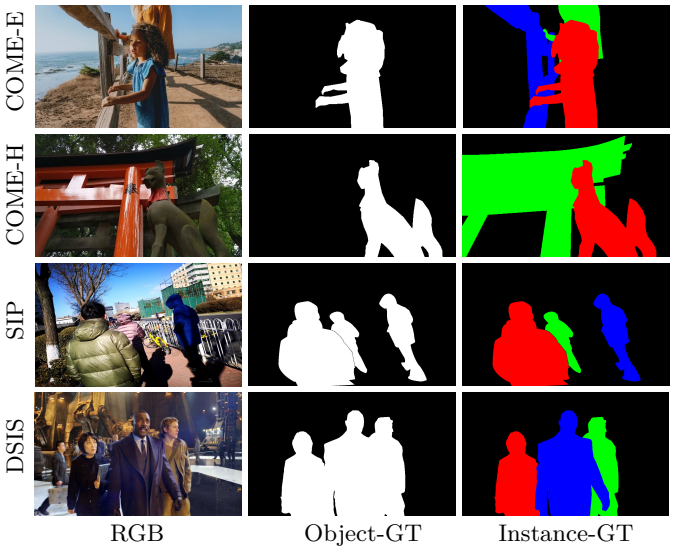


图 7. RGB-D 显著性数据集中显著区域级标签与实例级标签一致性对比。

者的匹配程度都低于 50%，这表明目标距离越近并不一定具有更高的显著性。从这个角度来看，对深度信息进行校准和过滤是至关重要的。在所有数据集中，DSIS 数据集的深度图与显著性区域标签的一致性是最低的，这也对跨模态模型提出了更大的挑战。本文希望所提出的 DSIS 数据集能够激励更多优秀 RGB-D SIS 模型的涌现，并设计更好的跨模态信息过滤与整合方法。

IV. *CalibNet*

A. 总体架构

如图 8所示，整体网络架构可以分为四个部分：(1) 两个分别用于提取多级 RGB 和深度特征的编码器；(2) 用于消除外观和结构信息之间的固有差异并过滤深度信息的深度相似性评估模块；(3) 用于对跨模态特征进行交互生成实例感知卷积核的卷积核分支；(4) 能够隐式校准 RGB 和深度共同特征并融合多尺度特征生成掩码特征的掩码分支。

B. 模型编码器

CalibNet 使用两个不共享参数的编码器来提取多尺度的 RGB 特征和深度特征。以 RGB 编码器为例，模型首先将输入图像 I 送入 ResNet-50 [53] 主干网络中，得到多尺度图像特征，记为 $\{C_i|i{=}2,3,4,5\}$。为了提升模型的全局感知能力并得到更深层的特征，本文在主干网络后使用了可变 Transformer [54] 进一步提取全局特征（$\{T_c^i|i{=}3,4,5\}$）。具体来说，C_3-C_5 被展平后拼接为一个序列作为 Transformer 的输入。为了进一步建立长期依赖关系，本文使用了原始的可变 Transformer 编码器，其中包括了一个多头可变自注意力模块，一个

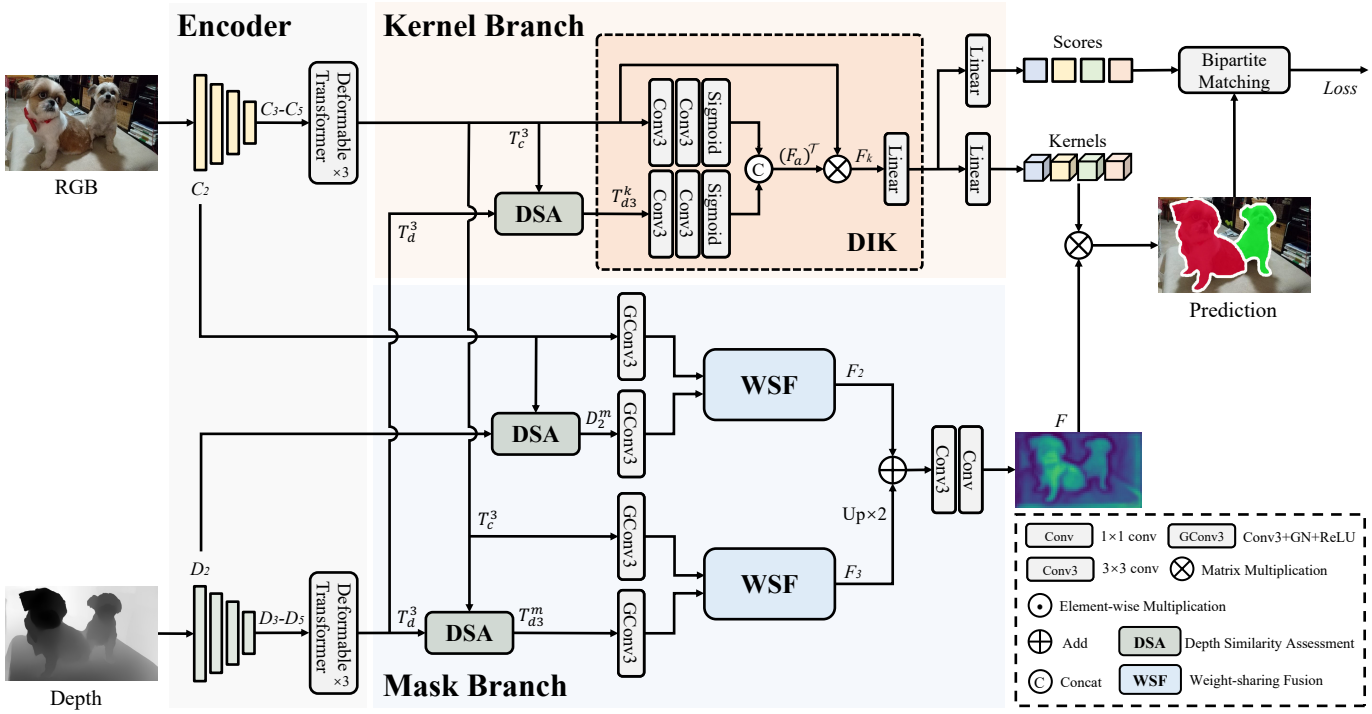


图 8. *CalibNet* 框架细节。*CalibNet* 包含两个并行的分支：卷积核分支和掩码分支。卷积核分支中，动态交互卷积核（DIK）模块（§ IV-C）用于 RGB 和深度特征的交互，以生成实例感知的卷积核；掩码分支中，本文提出的共享权重融合（WSF）模块（§ IV-D）能够在各个尺度上识别并校准跨模态信息，最后整合多尺度特征生成统一的掩码特征。深度相似性评估（DSA）模块（§ IV-E）被嵌入于两个分支中通过 RGB 和深度特征的相似性来校准深度信息。

层归一化操作和一个前馈神经网络（FFN）。RGB 编码器堆叠了三个可变 Transformer 层，并将最后一个层的输出序列重新恢复成全局特征 T_c^3-T_c^5。深度图分支使用了与 RGB 编码器具有相同结构的卷积主干网络，但二者之间不共享权重。深度分支中卷积主干网络输出的多尺度特征记为 $\{D_i|i{=}2,3,4,5\}$。深度分支中同样使用了三层可变 Transformer 层来生成更高级别的深度特征，记为 T_d^3-T_d^5。两个模态编码器输出的多尺度特征将送入实例卷积核分支和掩码特征分支以进一步融合。

C. 卷积核分支

卷积核分支旨在整合 RGB 和深度特征，以挖掘跨模态嵌入向量并生成一组实例感知的卷积核。为此，本文设计了动态交互卷积核（DIK）模块以利用双模态信息来促进跨模态特征交互，以提升实例感知特征的粒度。如图 8所示，RGB 特征 $T_c^3 \in \mathbb{R}^{h\times w\times c}$ 和经过 DSA（在 § IV-E中介绍）过滤的深度特征 $T_{d3}^k \in \mathbb{R}^{h\times w\times c}$ 作为 DIK 模块的输入。这里 h、w 和 c 分别表示特征图的高度、宽度以及通道数。与 [55] 保持一致，DIK 首先将两通道固定坐标拼接到原始特征图上，以此来增强特征的空间位置信息。然后两个模态的特征分别经过两个连续的 3×3 卷积层。为了优先考虑具有潜在实例感知能力的嵌入特征，在后一个实例感知的卷积操作中将两个模态特征的通道数从 $c+2$ 减小到目标显著实例卷积核个数 N，最后经过一个 Sigmoid 运算并将两个模态进行拼接，得到跨模态实例感知注意力图 $F_a \in \mathbb{R}^{h\times w\times N}$。该过程可以表述为

$$F_a = Cat[\mathcal{S}(\mathcal{C}_3(\mathcal{C}_3(T_c^3)));\mathcal{S}(\mathcal{C}_3(\mathcal{C}_3(T_{d3}^k)))], \qquad (1)$$

其中 $\mathcal{S}$ 表示 Sigmoid 函数，$\mathcal{C}_3$ 代表 3×3 卷积，$Cat[;]$ 代表按通道拼接操作。根据注意力权重的引导，$F_a \in \mathbb{R}^{h\times w\times N}$ 可以视为 Key 和 Value 用于与 RGB 特征 $T_c^3 \in \mathbb{R}^{h\times w\times c}$ 矩阵相乘，得到实例感知嵌入向量 $F_k \in \mathbb{R}^{N\times c}$。该操作可以描述为

$$F_k = (F_a)^{\mathcal{T}} \otimes T_c^3, \qquad (2)$$

其中，$\otimes$ 表示矩阵乘，$(\cdot)^{\mathcal{T}}$ 为矩阵转置操作。与其它使用交叉注意力操作获取卷积核的方法 [13], [14] 相比，本文的 DIK 模块更加轻量化，能够适应各种多模态输入。最后，给定 F_k，DIK 通过一个共享的线性层以及两个单独的线性层得到一组置信度分数 $S^{N\times 1}$ 和实例感知的卷积核 $K^{N\times c}$。

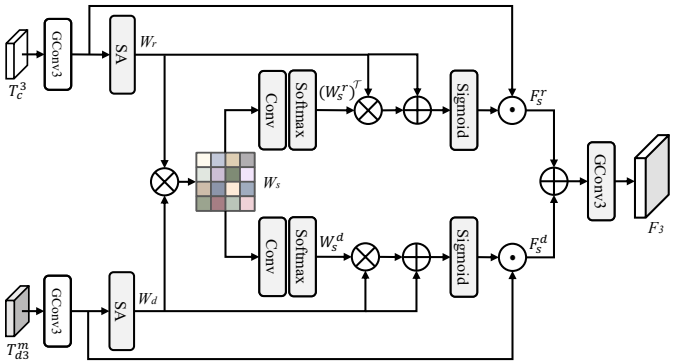


图 9. WSF 模块结构。其它尺寸（C_2, D_2^m）也采用了相同的操作。SA 表示生成空间权重的空间注意力机制 [56]。

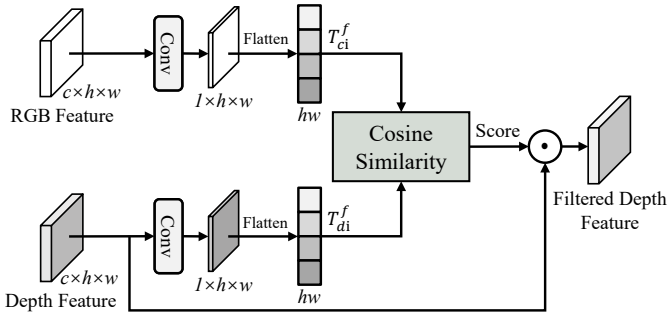


图 10. DSA 模块结构。

D. 掩码分支

平行于卷积核分支，掩码分支的作用是从跨模态线索中发现和校准特征息，并融合多尺度的低级别和高级别特征，最终生成统一的掩码特征。*CalibNet* 将来自 CNN 主干网络的浅层特征 C_2 与可变 Transformer 中的高级特征 T_c^3 一起作为输入掩码分支的输入，用于提供丰富的外观特征表示。与 C_2 和 T_c^3 相对应，来自深度分支的深度特征 D_2 和 T_d^3 也被用来增强特征中的结构信息、保持显著区域内部的一致性。模型将来自不同模态的相同尺度的特征分为同一组（如 $\{C_2, D_2\}$ $\{T_c^3, T_d^3\}$）。掩码分支的细节如图 8的底部所示。与卷积核分支相似，输入的深度特征 D_2 和 T_d^3 先经过 DSA 模块以获得过滤后的深度特征 D_2^m 和 T_{d3}^m。在经过一系列操作（3×3、组归一化、ReLU 激活函数）后，RGB 特征和深度特征被送入所提出的 WSF 模块，以有效地挖掘不同级别的跨模态互补特征来获取共享的特征表示。如图 9所示，WSF 模块旨在基于共享的亲和权重隐式地校准与整合跨模态特征，在两个模态之间建立更强的关联性，突出共同信息。为了实现上述目标，WSF 首先使用空间注意力机制（SA）[56] 为两个模态特征分别生成空间注意力权重 $W_r \in \mathbb{R}^{h\times w}$ 和 $W_d \in \mathbb{R}^{h\times w}$。以深度特征输入 T_{d3}^m 为例，SA 的计算过程可以表述为

$$W_d = \mathcal{S}(\mathcal{C}_7(Cat[\mathcal{A}(T_{d3}^m); \mathcal{M}(T_{d3}^m)]), \tag{3}$$

其中 $\mathcal{A}$ 表示平均池化操作，$\mathcal{M}$ 代表最大池化操作。同样的，D_2^m、C_2 和 T_c^3 也经过相同的操作来获取对应的空间注意力权重。为了更好地整合并校准跨模态特征，WSF 采用 W_r 和 W_d 一起生成共享的亲和权重 $W_s \in \mathbb{R}^{h\times h}$，计算过程为

$$W_s = W_r \otimes (W_d)^{\mathcal{T}}. \tag{4}$$

值得注意的是，相比于 [57]–[59]，这里的共享亲和矩阵 W_s 有着更小的尺寸 $h \times h$，能够避免计算像素-像素的亲和权重。此外，使用更小尺寸的亲和权重也能够减轻 WSF 中后续步骤的计算负担（见 § V-E）。随后，WSF 利用得到的共享权重 W_s，使用两个 1×1 卷积以及 softmax 函数分别生成特定模态的权重 $W_s^r \in \mathbb{R}^{h\times h}$ 和 $W_s^d \in \mathbb{R}^{h\times h}$。再将其分别与之前的空间权重 W_r 和 W_d 通过矩阵乘法以及残差连接的操作进行融合，以提升特定模态的空间注意力并保留原始模态特征的信息。最后使用矩阵点乘的方式对 RGB 特征以及深度特征进行重加权，得到重加权后的 RGB 特征 F_s^r 和深度特征 F_s^d。以输入特征 T_c^3 和 T_{d3}^m 为例，两个模态的重加权特征可以通过用如下公式得到

$$\begin{cases} F_s^r = \mathcal{G}_3(T_c^3) \odot \mathcal{S}(W_r + (W_s^r)^{\mathcal{T}} \otimes W_r) \\ F_s^d = \mathcal{G}_3(T_{d3}^m) \odot \mathcal{S}(W_d + W_s^d \otimes W_d) \end{cases} \tag{5}$$

其中 $\odot$ 代表元素点乘操作，$\mathcal{G}_3$ 代表由一个 3×3 卷积、一个组归一化以及一个 ReLU 函数组成的操作。最后将重加权后的 RGB 和深度特征（F_s^r 和 F_s^d）相加，得到融合后的原图 1/8 大小的跨模态特征 F_3。类似的，以 C_2 和 D_2^m 作为输入，使用 WSF 可以得到 1/4 大小的融合特征 F_2。整合不同尺度的融合特征对于生成共享掩码特征来说是至关重要的。因此，本文将 F_3 上采样至 1/4 大小并与 F_2 相加，然后使用一个 3×3 和一个 1×1 卷积得到最终统一的掩码特征 F。最后，实例感知的卷积核 $K \in \mathbb{R}^{N\times c}$ 以及掩码特征 $F \in \mathbb{R}^{c\times h\times w}$ 经过动态卷积操作最终得到显著实例掩码 $M \in \mathbb{R}^{N\times h\times w} = K \otimes F$。

E. 深度相似性评估模块 *DSA*

深度图中的结构信息与显著性并不是强烈相关的，低质量的深度图通常包含各类噪声信息，可能会干扰对显著实例的预测。为此，本文提出了深度相似性评估

（DSA）模块，通过计算相似性分数来衡量 RGB 和深度信息之间的关联，并以此动态调整深度特征的权重。在 *CalibNet* 中，DSA 模块被嵌入于卷积核分支以及掩码分支中来校准深度特征。DSA 模块的结构如图 10所示。两个模态的特征对称地经过一个 1×1 卷积生成各自的注意力图。然后，RGB 和深度的注意力图被展平，并计算两个特征向量之间的相似度。值得注意的是，为了使得该相似度分数对于方向性和空间性更加敏感，DSA 采用了余弦相似度作为相似性度量。对于展平后的权重图向量 T_{ci}^f 和 T_{di}^f，相似性分数 V_s 的计算方式为

$$V_s = 1 - (1 - \frac{T_{di}^f \odot T_{ci}^f}{\left\|T_{di}^f\right\| \left\|T_{ci}^f\right\|})/2. \tag{6}$$

相似度分数 V_s 的取值范围为 [0, 1]，其值越高则表明不同模态之间的特征一致性越高。最后，DSA 将相似性分数与输入深度特征相乘，来控制深度信息在特征融合过程中的占比。

F. 损失函数

为了进行端到端训练，提出的 *CalibNet* 在训练时使用二分匹配对置信度分数以及示例掩码进行分配。与 DETR [14] 保持一致，本文使用匈牙利算法来优化真值和预测结果之间的匹配。总体的损失函数如下所示：

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_{mask} \mathcal{L}_{mask} + \lambda_{obj} \mathcal{L}_{obj} + \lambda_{bin} \mathcal{L}_{bin}, \tag{7}$$

其中 $\mathcal{L}_c$ 为使用置信度分数区分显著实例的显著性定位损失。$\mathcal{L}_c$ 采用的是 focal loss [71]。受 [38] 的启发，本文也采用了二元交叉熵损失来优化 IoU 感知的目标分数 $\mathcal{L}_{obj}$。此外，本文还采用了辅助损失 $\mathcal{L}_{bin}$ 对于显著区域的预测结果进行深度监督，其公式表述为

$$\begin{aligned} \mathcal{L}_{bin} = & \sum_{r \in \{C_2, T_3\}} \lambda_r \mathcal{L}_{bce}(\mathcal{P}_r, \mathcal{R}_o) + \\ & \sum_{d \in \{D_2, D_3\}} \lambda_d \mathcal{L}_{bce}(\mathcal{P}_d, \mathcal{R}_o), \end{aligned} \tag{8}$$

其中 $\mathcal{L}_{bce}$ 是二元交叉熵损失，$\mathcal{R}_o$ 表示由示例标签得到的区域级真值。$\mathcal{P}_r$ 和 $\mathcal{P}_d$ 由表示编码器输出的 RGB 和深度特征经过一个 1×1 卷积核 Sigmoid 函数生成的区域级预测结果。对应的权重系数 λ_r 和 λ_d 分别设置为 0.6 和 0.4。此外，λ_c、λ_{mask}、λ_{obj} 和 λ_{bin} 分别设置为 2，1，1，1 来平衡总体的损失函数。

V. 实验

A. 数据集

作为一项新的挑战，RGB-D SIS 采用了 RGB-D SOD 数据集中的实例级标签，包括 COME15K [15] 和 SIP [16]。COME15K 训练集包括 8,025 张带有实例级标签的图像。COME15K 的测试集按难度分为两个不同的子集，分别为包含 4,600 张一般难度样本的 COME15K-E 以及包含 3,000 张困难样本的 COME15K-H。在此之前，Fan 等人 [16] 发布了一个带有实例级标签的 RGB-D SOD 数据集，名为 SIP。其中包含了 929 张高像素图片用作测试集。为了全面评估 RGB-D SIS 方法的性能，本文发布了一个新的 RGB-D SIS 测试集 DSIS，包含了共计 1,940 组样本，并带有高质量的实例级标签。DSIS 数据集的更多细节可以参考 § III. 本文采用 COME15K 中的实例级训练集来训练 *CalibNet*，并在所有数据集的测试集上对预测结果进行评估。

B. 评价指标

与先前 SIS 方法 [40], [43] 保持一致，本文同样采用 COCO 样式的评价指标来评估 RGB-D SIS 模型的性能。其中，AP（average precision）分数可以用于反映被测试模型的综合性能。AP_{50} 和 AP_{70} 可以在特定掩码 IoU 阈值下，更细致地评估显著实例的预测结果。

C. 实现细节

CalibNet 基于 Detectron2 [72] 框架实现，在一张 RTX 3090 GPU 上以批量大小为 4 进行训练和测试。在训练阶段，所有的模型都以 2.5e-5 为初始学习率，使用 AdamW 优化器 [73] 训练了 50 次迭代，学习率会在第 35 和第 45 次迭代除以 10，并设置了 0.0001 的权重衰减。除非特别声明，本文默认采用在 ImageNet [74] 上预训练的 ResNet-50 [53] 权重初始化 CNN 主干网络，网络中的其他参数都是随机初始化的。输入的 RGB 和深度图都根据原始长宽比调整为短边为 320 像素，长边最多为 480 像素。本文的实验中还采用了随机翻折和尺度抖动进行数据增强。*CalibNet* 中默认卷积核数量 N 被设置为 50，RGB 和深度编码器中的 Transformer 层数量均设置为 3。

D. 与先进方法的对比

据我们所知，*CalibNet* 是第一个 RGB-D SIS 模型。为了更全面地比较，本文从其它相关

表 I
RGB-D SIS 任务上与 19 个相关任务代表性方法的定量对比。

Task	Method	Volume	COME15K-E [15]			COME15K-H [15]			DSIS			SIP [16]		
			AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}
SIS	S4Net [11]	CVPR19	43.7	68.0	52.5	37.1	60.9	43.2	58.3	81.9	71.8	49.6	76.0	63.7
	RDPNet [40]	TIP21	49.8	72.2	59.5	42.1	65.2	49.7	66.1	87.2	80.1	59.0	80.1	74.1
	OQTR [43]	TMM22	48.7	71.1	58.3	42.7	65.9	50.5	63.1	85.9	77.0	59.9	83.1	76.3
CIS	OSFormer [60]	ECCV22	53.0	71.9	61.3	45.8	66.3	52.4	67.9	86.3	78.3	63.2	80.5	74.6
GIS	Mask R-CNN [10]	ICCV17	48.8	71.2	58.6	42.2	65.7	50.8	65.6	86.8	80.2	57.9	79.8	73.3
	MS R-CNN [61]	CVPR19	49.7	70.4	58.9	42.3	63.8	50.0	66.8	87.3	80.3	60.0	79.8	73.3
	Cascade R-CNN [62]	TPAMI19	49.4	69.8	58.7	41.9	64.3	49.8	66.4	87.3	79.9	59.1	79.1	73.6
	CenterMask [63]	CVPR20	49.6	71.6	59.0	42.5	65.2	51.0	65.7	87.6	79.7	57.6	79.8	72.3
	HTC [64]	CVPR19	51.4	73.7	61.0	44.1	68.0	52.1	67.5	87.6	81.1	60.0	81.4	74.9
	YOLACT [65]	ICCV19	48.1	70.7	56.2	41.4	66.0	48.5	62.4	84.1	73.3	62.0	82.3	74.7
	BlendMask [66]	CVPR20	48.1	70.6	56.9	41.0	64.8	48.5	65.5	86.9	78.0	55.5	77.3	69.7
	CondInst [67]	ECCV20	49.6	72.0	59.5	42.8	66.4	51.0	65.1	86.8	79.1	58.6	79.4	73.3
	SOLOv2 [12]	TPAMI21	51.1	71.6	59.9	45.1	66.3	52.8	67.4	87.0	80.4	63.4	80.7	74.8
	QueryInst [68]	CVPR21	51.5	73.1	61.1	43.9	67.5	51.8	67.9	87.5	80.6	61.3	81.0	75.1
	SOTR [69]	ICCV21	50.7	70.4	59.0	43.5	65.1	50.3	68.2	86.9	79.3	61.6	78.4	73.0
	Mask Transfiner [70]	CVPR22	48.7	71.0	56.3	40.7	64.5	47.2	67.5	87.6	80.4	57.8	78.9	70.3
	SparseInst [38]	CVPR22	51.3	71.9	58.9	43.1	65.1	48.5	65.0	85.0	75.7	62.8	81.3	75.3
	Mask2Former [13]	CVPR22	51.5	67.3	57.6	44.1	62.4	49.4	68.3	85.1	76.7	66.6	79.3	75.1
RGB-D GIS	RRL [9]	TIP20	52.9	71.8	61.8	45.6	66.1	53.6	66.1	87.1	79.6	61.9	79.1	74.6
RGB-D SIS	***CalibNet* (Ours)**	-	**58.0**	**75.8**	**65.6**	**50.7**	**70.4**	**57.3**	**69.3**	**87.8**	**81.6**	**72.1**	**86.6**	**82.9**

任务中选取了 19 个具有代表性的模型，包括：显著实例分割模型 S4Net [11]、RDPNet [40] 和 OQTR [43]；伪装实例分割（CIS）模型 OSFormer [60]；通用实例分割（GIS）模型 Mask R-CNN [10]、MS R-CNN [61]、Cascade R-CNN [62]、CenterMask [63]、HTC [64]、YOLACT [65]、BlendMask [66]、CondInst [67]、SOLOv2 [12]、QueryInst [68]、SOTR [69]、Mask Transfiner [70]、SparseInst [38] 以及 Mask2Former [13]；RGB-D 通用实例分割模型 RRL [9]。为了进一步比较，本文还对上述模型中性能前三的模型加入深度模态，以在 RGB-D SIS 任务中与提出的模型进行更直接的对比。值得注意的是，实验过程中本文统一使用了对比方法的官方代码在 COME15K 数据集上进行训练，并在 COME15K、SIP 和本文的 DSIS 测试集上评估性能。上述模型同样采用了在 ImageNet [74] 上预训练的 ResNet-50 [53] 作为主干网络。

定量比较。所有对比模型的预测结果如表 I所示。本文提出的 *CalibNet* 在三个数据集上都以较大的优势超过了对比方法，比如在 COME15K-E 测试集上有 5.0% AP 的提升。即使在更具挑战性的 COME15K-H 测试集上，*CalibNet* 仍旧达到了 50.7% AP 以及 70.4% AP_{50}。这证明了本文方法在多种复杂场景下都具有良好的鲁棒性。此外，从表中可以观察到 RGB-D 跨模态模型在 DSIS 测试集上的性能优势并不十分明显，比如 RRL [9] 和 *CalibNet* 在 DSIS 上的 AP 分别为 66.1% 和 69.3%。这个现象是因为 DSIS 数据集中显著区域和深度图数据具有更低的一致性（参考图 5(c)），从而在利用深度信息时具有更大的难度。相比之下，大部分仅

表 II
与融合了深度模态后的实例级 SOTA 模型的性能比较。

Method	COME15K-E [15]			COME15K-H [15]			Params	FPS
	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}		
OSFormer*	50.9	71.0	58.6	43.3	65.2	50.2	78.9M	28.1
Mask2Former*	55.9	72.7	62.4	48.3	66.9	54.1	**68.3**M	28.0
QueryInst*	51.5	73.4	61.4	44.6	67.7	52.8	199.8M	16.9
***CalibNet* (Ours)**	**58.0**	**75.8**	**65.6**	**50.7**	**70.4**	**57.3**	78.9M	**35.9**

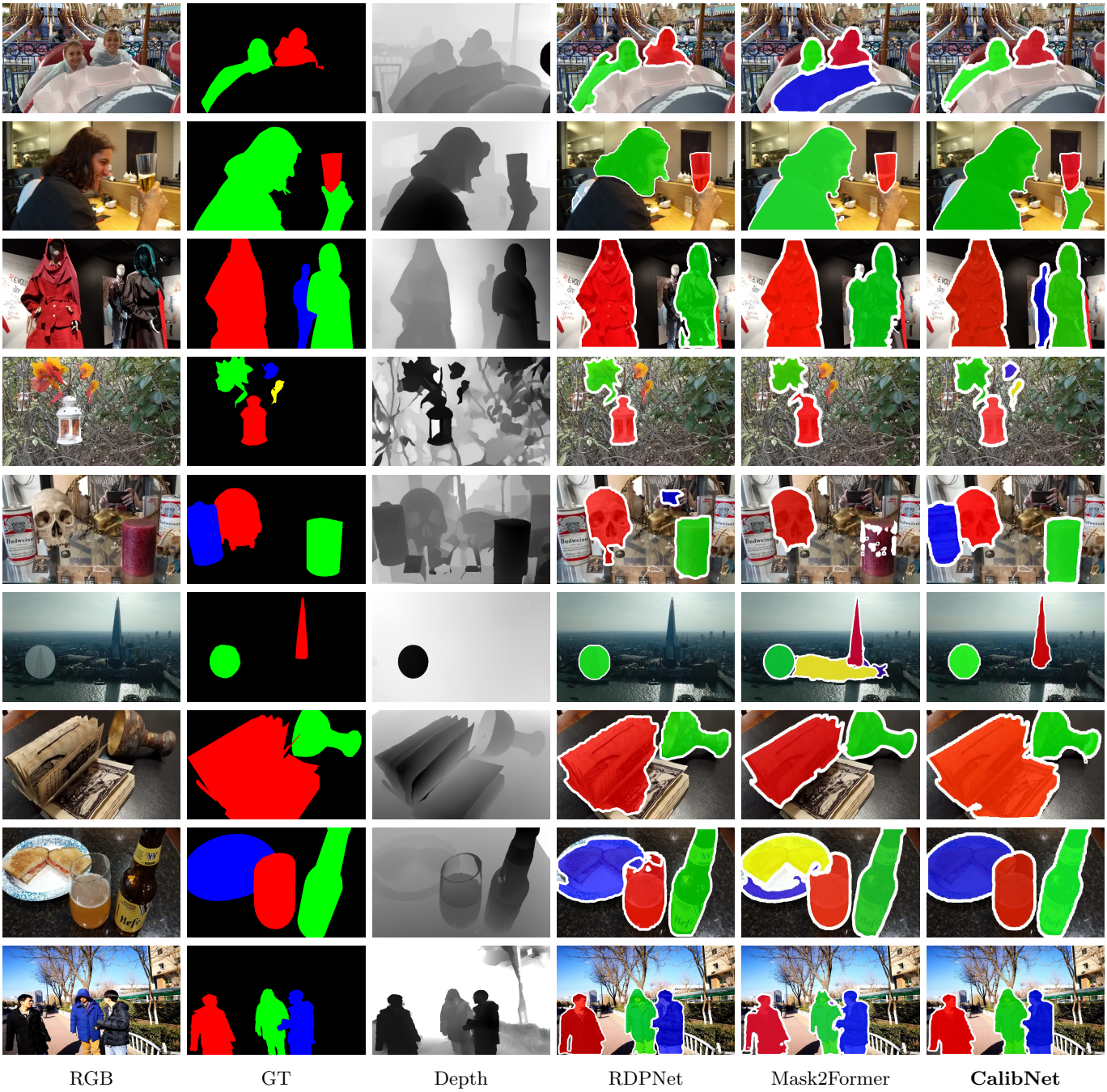


图 11. Visual comparison of *CalibNet* with representative models on all four RGB-D SIS test sets.

使用 RGB 模态的 GIS 模型则不需要考虑深度图与显著性之间的一致性问题。为了进一步比较，本文选取了对比方法中性能最好的三个以加入深度模态并重新进行训练（表 II 中的 OSFormer*、Mask2Former* 以及 QueryInst*）。对于每个模型，本文平行于原有的 RGB 主干网络嵌入了一个单独的主干网络来提取多尺度深度特征，然后再将两个模态的多尺度特征按通道进行拼接。随后使用一个 1×1 卷积和一个组归一化来得到跨模态融合后的特征，同时将通道数减半。原始架构中其余操作均保持不变。如表 II所示，融合深度数据后，Mask2Former* 的 AP 提升了 4.4%，而 OSFormer* 的性能却略有下降。整体而言，本文的方法相比于其它融合深度模态的模型仍旧取得了领先的性能。这表明了本文所提出 DIK、WSF 等模块能够有效地整合与利用深度模态中有用的信息。此外，在多模态的架构下，我们模型推理的 fps 为 35.9，能够满足实时推理的需求。上述结果同时也表明深度模态有助于提高实例级显著性检测模型的性能，这也与 [25] 中的结论相吻合。

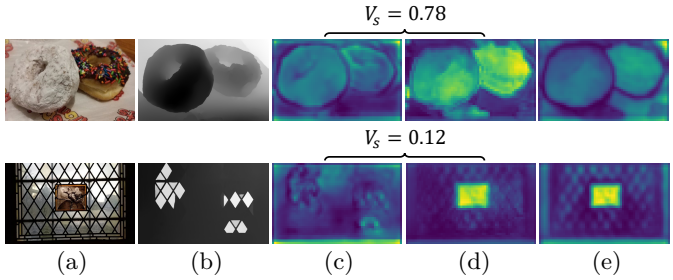


图 12. 掩码分支中注意力图的可视化。顶部是一个高像素的样例，底部是一个低相似度的样例。(a) 输入图片；(b) 深度图；(c) 输入 DSA 的深度特征；(d) 输入 DSA 的 RGB 特征；(e) 生成的掩码特征。

表 III
DIK 中 RGB 和深度特征融合方式的消融实验。

DIK	COME15K-E			COME15K-H		
	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}
早期融合	57.2	75.7	65.0	48.7	69.6	54.9
后期融合	**58.0**	**75.8**	**65.6**	**50.7**	**70.4**	**57.3**

定性比较。本文进一步将 *CalibNet* 与 RDPNet [40] 和 Mask2former [13] 在四个测试集上的结果进行了对比，并展示了一些典型的可视化结果。如图 11所示，本文提出的 *CalibNet* 能够在多种场景下更准确地识别并分割大部分视觉上显著的实例目标，同时抑制杂乱的背景与不相关的目标。例如，图 11中第一行的样本中包含了两个女孩和多种游乐设施，*CalibNet* 能够准确地区分她们并抑制图像前部的游乐设施。即使是在图 11中第四、第五行的一些带有较多重叠物体的复杂背景下，我们的模型仍能够准确地检测到显著实例。在面对光照变化的场景时，尤其是目标尺寸相对较小的时候（图中的第六行），本文中的方法也同样能够处理。更重要的是，作为实例级的 RGB-D 模型，即使是深度图中的结构信息与显著区域不对齐（图中第一行和第六行），*CalibNet* 仍可以有效地进行处理。这也证明了提出的 WSF、DSA 等模块在校准跨模态特征时的有效性。与 RDPNet 以及 Mask2former 的可视化结果相比，*CalibNet* 可以准确地捕捉实例的边缘区域，这得益于深度信息带来的贡献以及对深度信息有效的校准，无论在多目标或出现遮挡时都能得到完整的实例掩码。

E. 消融实验

为了评估提出模型的有效性，本文进行了一系列消融实验来验证 *CalibNet* 中的每个部分。

表 IV
在 COME15K-E 测试集上对不同的共享亲和权重 W_s 设计方式比较。

Shared Weight	AP	AP_{50}	AP_{70}	Params	FLOPs	FPS
Non-local ($hw \times hw$)	**57.4**	**75.4**	**65.0**	75.9M	64.6G	29.4
Ours ($h \times h$)	57.0	75.1	64.5	75.9M	**54.0**G	**39.6**

表 V
CalibNet 中共享 DSA 模块的消融实验。

DSA	COME15K-E			COME15K-H		
	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}
Shared	57.0	75.6	64.2	50.3	70.4	56.7
w/o Shared	**58.0**	**75.8**	**65.6**	**50.7**	**70.4**	**57.3**

DIK 中的模态融合方式。DIK 模块中，本文尝试将 RGB 和深度特征融合来生成实例卷积核。除了使用在 Sigmoid 函数后融合两个模态特征这种默认的方式（后期融合），我们也尝试了首先将两个模态特征进行拼接，然后共同进行后续操作（早期融合）。表 III表明后期融合取得了更高的性能，这是由于后期融合的设计可以在生成实例感知卷积核时分别利用结构信息和外观特征，以覆盖更多潜在候选实例嵌入。

WSF 中的共享权重。WSF 模块设计的核心是共享亲和权重 $W_s \in \mathbb{R}^{h \times h}$，它能够在 RGB 和深度模态之间建立强关联性，挖掘其中的共同信息。为了证明其高效性，本文尝试采用非局部的方式 [57], [58] 来生成 WSF 中的共享亲和权重，记为全局共享权重 $W_s^g \in \mathbb{R}^{hw \times hw}$。具体来说，RGB 和深度模态的空间注意力权重 $W_r \in \mathbb{R}^{h \times w}$ 和 $W_d \in \mathbb{R}^{h \times w}$ 首先被展平为 $L_r \in \mathbb{R}^{1 \times hw}$ 和 $L_d \in \mathbb{R}^{1 \times hw}$。然后 L_r 和 L_d 也同样使用矩阵乘法来计算 W_s^g。为了便于进行比较，实验仅使用单尺度特征（T_c^3 和 T_d^3）融合进掩码特征之中。如表 IV所示，使用非局部共享权重的模型 FLOPS 显著增加，但准确率的提升却相对有限。当加入浅层特征（C_2 和 D_2）后，FLOPS 还会进一步增加到 $103G$。除此之外，全尺寸的权重图还会提升后续操作的计算量，导致在推理时的 FPS 显著下降。总之，默认的共享亲和权重 W_s 在保证性能的同时让 *CalibNet* 更加轻量化。

DSA 的共享设置。*CalibNet* 在 DIK 和 WSF 之前嵌入了多个 DSA 模块来过滤深度特征以实现更好的特征一致性。在表 V中，我们消融了掩码分支和卷积核分支是否共享 DSA 模块。结果表明，使用共享 DSA

表 VI
在 COME15K-E 测试集上对掩码分支中多尺度特征融合方式的对比。

Combinations	AP	AP_{50}	AP_{70}	Params	FLOPs	FPS
(T_c^3,T_d^3)	57.0	75.1	64.5	**75.9**M	**54.0**G	**39.6**
(C_2,D_2); (T_c^3,T_d^3)	**58.0**	**75.8**	**65.6**	78.9M	90.2G	35.9
(C_2,D_2); (T_c^3,T_d^3); (T_c^4,T_d^4)	57.3	75.6	65.0	81.9M	91.8G	35.2
(T_c^3,T_d^3); (T_c^4,T_d^4); (T_c^5,T_d^5)	57.1	75.4	64.7	81.9M	56.0G	35.8
(C_2,D_2); (T_c^3,T_d^3); (T_c^4,T_d^4); (T_c^5,T_d^5)	57.5	**75.8**	64.8	84.8M	92.2G	34.3

表 VII
不同实例感知的卷积核数量的消融实验。

N	COME15K-E			COME15K-H			Params	GFLOPs
	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}		
10	56.5	75.0	64.0	48.5	69.2	54.3	**78.7**M	**89.5**
20	57.2	75.2	64.6	50.0	70.3	56.1	78.8M	89.7
50	58.0	75.8	**65.6**	**50.7**	70.4	**57.3**	78.9M	90.2
100	**58.1**	**76.3**	**65.6**	50.3	**70.9**	57.2	79.2M	91.2

表 VIII
CalibNet 在 RGB 和深度分支中使用不同 Transformer 层数的比较。

Number		COME15K-E			COME15K-H			Params
RGB	Depth	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}	
3	0	57.2	75.4	64.9	49.6	70.5	56.4	76.7M
3	1	57.1	75.7	65.0	49.0	69.9	55.3	77.4M
3	3	58.0	75.8	65.6	**50.7**	70.4	**57.3**	78.9M
6	1	55.4	74.6	63.2	47.2	69.0	52.9	79.6M
6	3	**58.2**	**76.3**	**65.8**	50.6	**70.9**	57.0	81.1M
6	6	58.0	76.1	65.3	50.4	70.9	56.8	83.3M

表 IX
WSF 与其它融合策略的对比。

Interactions	COME15K-N			COME15K-D		
	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}
Cascade fusion	57.1	75.4	64.7	49.8	69.9	56.1
RMFF [75]	57.6	75.6	65.1	49.9	70.1	55.9
WSF (Ours)	**58.0**	**75.8**	**65.6**	**50.7**	**70.4**	**57.3**

模块的 *CalibNet* 性能相对较差，这是因为掩码特征分支和实例卷积核分支对 RGB 和深度特征离散化的需求不同。因此，文中框架最终在不同分支中采用非参数共享的 DSA 模块。图 12可视化了 DSA 中 RGB 和深度特征的一些典型样例以及其对应的相似性分数。可以看出两个模态特征之间的视觉差异与量化分数是一致的。

掩码分支中融合的特征尺度。 *CalibNet* 中的掩码分支使用 WSF 融合跨模态特征的同时，还需要将 RGB 和深度编码器的多尺度特征进行融合，得到统一的掩码特征。因此，本文在掩码分支中使用不同层级的特征进行了消融实验，来评估模型的有效性。不同的特征层级组合如表 VI所示。前两行结果能够说明浅层特征的重要性。在此基础上，加入更多层的特征不会带来显著的性能提升，反而会引入额外的参数量。考虑到准确率与效率的平衡，本文最终采用了 CNN 主干网络输出的第二层特征（C_2 和 D_2）Transformer 输出的第三层特征（T_c^3 和 T_d^3）进行多尺度融合。掩码分支最终生成的掩码特征也在图 12最后一列进行了展示。

卷积核数量。合适数量的实例感知卷积核能够在有效定位潜在显著实例的同时平衡模型的准确率与效率。为了探究不同卷积核数量的影响，本文对卷积核数量 N（10、20、50 和 100）进行了消融实验。如表 VII所示，随着卷积核数量的增加，*CalibNet* 的性能也会随之增强。当 N 大于 50 时，准确率的提升较为有限，甚至在 COME15K-H 测试集上的 AP 略有下降。考虑到模型的准确率与运行效率，最终 *CalibNet* 的卷积核数量被设置为 50。

Transformer 层的数量。 *CalibNet* 在 RGB 和深度编码器中使用了可变 Transformer 层来提高模型对全局信息的获取能力。为了优化模型的性能，本文对于 RGB 和深度编码器中不同 Transformer 层数的组合进行了消融实验。如表 VIII所示，第一行结果表明仅在 RGB 分支中使用 Transformer 层并不合适。当 RGB 和深度分支中都采用 3 层 Transformer 的时候，CalibNet 能够在 COME15K-E 和 COME15K-H 测试集 [15] 上都取得不错的效果。虽然在 RGB 编码器中使用 6 层的可变 Transformer 层能够进一步提升在 COME15K-E 测试集上的性能，但额外的模型参数量也会对多模态架构造成负面影响。因此，*CalibNet* 中两个模态编码器中 Transformer 层数都被设置为 3。

与其它融合方式的对比。表 IX进一步将提出的 WSF 与其它特征交互方式进行了对比，以证明 WSF 的有效性。第一行代表使用级联融合以及通道压缩操作来替换 WSF 中的共享亲和权重。第二行中的 RMFF [75] 模块可以通过计算两个模态中间层特征的集合、共享以及差别信息来识别其中的冗余与互补关系。相比而言，本文所提出的 WSF 模块使用了空间注意力权重来生成

表 X
目标分数损失 $\mathcal{L}_{obj}$ 以及辅助损失 $\mathcal{L}_{bin}$ 作用的消融实验。

Loss Items		COME15K-E			COME15K-H		
$\mathcal{L}_{obj}$	$\mathcal{L}_{bin}$	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}
	✓	56.7	75.4	64.7	49.7	70.5	56.4
✓		57.8	**76.1**	65.4	50.2	**70.9**	57.1
✓	✓	**58.0**	75.8	**65.6**	**50.7**	70.4	**57.3**

表 XI
在 COME15K-E 测试集上对 CALIBNET 各个模块作用的消融实验。

Input	DIK	WSF	DSA	AP	AP_{50}	AP_{70}
RGB				52.9	72.8	60.5
RGB-D				55.2	74.1	62.9
	✓			56.5	75.3	64.0
		✓		56.7	75.1	64.1
			✓	55.8	75.4	64.3
	✓		✓	56.9	75.3	64.9
		✓	✓	57.3	75.6	64.7
	✓	✓		57.5	75.6	65.1
	✓	✓	✓	**58.0**	**75.8**	**65.6**

表 XII
使用不同主干网络时 CALIBNET 的性能比较。

Backbones	COME15K-E			COME15K-H		
	AP	AP_{50}	AP_{70}	AP	AP_{50}	AP_{70}
ResNet-50 [53]	58.0	75.8	65.6	50.7	70.4	57.3
ResNet-101 [53]	58.5	76.4	66.0	51.5	71.8	58.0
Swin-T [76]	60.0	77.9	67.5	52.6	72.9	58.6
PVTv2-B2-Li [77]	60.7	78.1	67.8	53.7	72.9	59.6
P2T-Large [78]	**61.8**	**78.8**	**68.6**	**54.4**	**73.5**	**60.3**

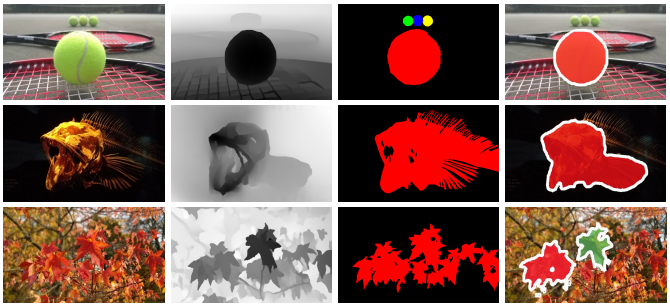


图 13. *CalibNet* 的典型预测不理想样例。

共享亲和图，这样能够更好地利用深度图中的结构信息，对跨模态信息进行校准和融合。表 IX表明 WSF 模块在所有的融合策略中取得了较好的性能，这也证明了 WSF 模块的先进性。

损失项的作用。 *CalibNet* 在训练时使用了 IoU 感知的目标分数损失来区分实例，并且使用了区域级的辅助监督。本文进行了消融实验来探究这两个损失项对模型性能的贡献。如表 X所示，目标分数损失以及辅助损失都对模型的性能起到了积极的作用，在 COME15K-E 测试集上分别带来了 1.3% 和 0.2% 的 AP 提升。

各个模块的作用。为了优化模型的性能，本文进行了消融实验来验证 *CalibNet* 中各个模块的作用，包括动态交互卷积核（DIK）模块、共享权重融合（WSF）模块以及深度相似性评估（DSA）模块。结果如表 XI所示，其中 RGB 基线模型代表不使用任何本文所提出的模块，仅使用 RGB 模态来生成实例感知的卷积核以及掩码特征。在使用 RGB-D 输入的情况下，表中第二行简单地将 RGB 和深度特征相加，然后使用 1×1 卷积进行融合。可以看到，在引入了深度模态之后，模型的 AP 性能提升了 2.3%，这得益于所提出的跨模态双分支架构。表中的接下来三行表明，通过有效地融合深度特征，每个模块都对模型的性能起到了积极的作用，尤其是 DIK 和 WSF。在使用了所有模块时，*CalibNet* 取得了最好的性能。

不同主干网络的性能。为了挖掘模型的性能潜力，本文采用了多个在 ImageNet [74] 上预训练的主干网络权重来训练 *CalibNet*，包括 ResNet-50 [53]，ResNet-101 [53]，Swin-T [76]，PVTv2-B2-Li [77] 以及 P2T-Large [78]。如表 XII所示，*CalibNet* 在默认的 ResNet-50 主干网络下取得了 58.0% 的 AP。在使用更强的主干网络时（如 P2T-Large），本文的模型在 COME15K-E 测试集上的 AP 达到了 61.8%，说明我们的模型性能还有着很大的潜力。

F. 局限性

本节展示了使用 ResNet-50 [53] 的 *CalibNet* 在 COME15K 测试集上的一些不理想样例，如图 13所示。图中第一行表明 CalibNet 在图像中同时出现大小目标的时候有可能会忽略对于小目标的预测，这可能是由于深度图信息在其中起到了负面的影响，导致生成的实例卷积核忽略了小目标的实例嵌入。中间的样本表明在处理一些低对比度的场景或是边缘复杂的目标时，可能会得到不理想的结果。这个问题可以通过引入边缘先验信息或嵌入高亮边缘线索来缓解。此外，如果场景中背景过度重叠（见最下方样本），*CalibNet* 有可能会对显著

实例过度分割。在今后的研究中这些问题需要被进一步解决，以处理日益复杂的生活场景，提升模型从重叠复杂背景中理解和区分显著实例的能力。

VI. 结论

本文针对 RGB-D 显著实例分割任务提出了一个双分支跨模态校准模型 *CalibNet*，所提出的模型包含了平行的两个分支：卷积核分支使用动态交互卷积核（DIK）模块有效地融合 RGB 和深度特征生成实例感知的卷积核；掩码分支使用共享权重的融合（WSF）模块整合多尺度跨模态特征生成统一的掩码特征。此外，深度相似性评估（DSA）模块也被嵌入于两个分支中用于抑制深度信息带来的噪声。大量的实验结果证明了 *CalibNet* 先进的性能以及在 SIS 任务之上引入深度图所带来的好处。另外，本文的作者们还构建了一个新的 RGB-D SIS 数据集，名为 DSIS，来对相关模型进行更全面的评估。同时，该任务在促进其他多模态检测任务发展方面也具有巨大的潜力。

参考文献

[1] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *CVMJ*, vol. 7, no. 1, pp. 37–69, 2021.

[2] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *IEEE ICCV*, 2021, pp. 4722–4732.

[3] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, "Calibrated rgb-d salient object detection," in *IEEE CVPR*, 2021, pp. 9471–9481.

[4] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving rgb-d saliency detection," in *IEEE ICCV*, 2021, pp. 4681–4691.

[5] H. Wen, C. Yan, X. Zhou, R. Cong, Y. Sun, B. Zheng, J. Zhang, Y. Bao, and G. Ding, "Dynamic selective network for rgb-d salient object detection," *IEEE TIP*, vol. 30, pp. 9179–9192, 2021.

[6] K. Chen, L. Hong, H. Xu, Z. Li, and D.-Y. Yeung, "Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving," in *IEEE ICCV*, 2021, pp. 7546–7554.

[7] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *IEEE CVPR*, 2021, pp. 14 009–14 018.

[8] Y. Jin, Y. Yu, C. Chen, Z. Zhao, P.-A. Heng, and D. Stoyanov, "Exploring intra-and inter-video relation for surgical semantic scene segmentation," *IEEE TMI*, 2022.

[9] Z. Xu, S. Liu, J. Shi, and C. Lu, "Outdoor rgbd instance segmentation with residual regretting learning," *IEEE TIP*, vol. 29, pp. 5301–5309, 2020.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE ICCV*, 2017, pp. 2961–2969.

[11] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-M. Hu, "S4net: Single stage salient-instance segmentation," in *IEEE CVPR*, 2019, pp. 6103–6112.

[12] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Solo: A simple framework for instance segmentation," *IEEE TPAMI*, 2021.

[13] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *IEEE CVPR*, 2022, pp. 1290–1299.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020, pp. 213–229.

[15] J. Zhang, D.-P. Fan, Y. Dai, X. Yu, Y. Zhong, N. Barnes, and L. Shao, "Rgb-d saliency detection via cascaded mutual information minimization," in *IEEE ICCV*, 2021, pp. 4338–4347.

[16] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE TNNLS*, vol. 32, no. 5, pp. 2075–2089, 2020.

[17] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE TIP*, vol. 26, no. 5, pp. 2274–2285, 2017.

[18] W. Zhang, Y. Jiang, K. Fu, and Q. Zhao, "Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection," in *ICME*, 2021, pp. 1–6.

[19] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation," *IEEE TIP*, vol. 27, no. 2, pp. 568–579, 2017.

[20] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection," *IEEE TIP*, vol. 30, pp. 7012–7024, 2020.

[21] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes, "Uncertainty inspired rgb-d saliency detection," *IEEE TPAMI*, 2021.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015, pp. 3431–3440.

[23] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE TIP*, vol. 26, no. 9, pp. 4204–4216, 2017.

[24] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time rgb-d salient object detection," in *ECCV*, 2020, pp. 646–662.

[25] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth." in *BMVC*, 2013, pp. 1–11.

[26] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion," *IEEE TCYB*, vol. 48, no. 11, pp. 3171–3183, 2017.

[27] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection," in *IEEE CVPR*, 2020, pp. 9060–9069.

[28] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *IEEE CVPR*, 2018, pp. 3051–3060.

[29] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE CVPR*, 2019, pp. 3927–3936.

[30] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *IEEE CVPR*, 2020, pp. 13 756–13 765.

[31] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *IEEE CVPR*, 2020, pp. 3472–3481.

[32] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate rgb-d salient object detection via collaborative learning," in *ECCV*, 2020, pp. 52–69.

[33] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *ECCV*, 2020, pp. 275–292.

[34] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for rgb-d salient object detection," in *ECCV*, 2020, pp. 665–681.

[35] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *IEEE ICCV*, 2019, pp. 7254–7263.

[36] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for rgb-d salient object detection and beyond," *IEEE TPAMI*, 2021.

[37] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, and M.-M. Cheng, "Mobilesal: Extremely efficient rgb-d salient object detection," *IEEE TPAMI*, 2021.

[38] T. Cheng, X. Wang, S. Chen, W. Zhang, Q. Zhang, C. Huang, Z. Zhang, and W. Liu, "Sparse instance activation for real-time instance segmentation," in *IEEE CVPR*, 2022, pp. 4433–4442.

[39] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE CVPR*, 2017, pp. 2386–2395.

[40] Y.-H. Wu, Y. Liu, L. Zhang, W. Gao, and M.-M. Cheng, "Regularized densely-connected pyramid network for salient instance segmentation," *IEEE TIP*, vol. 30, pp. 3897–3907, 2021.

[41] N. Liu, W. Zhao, L. Shao, and J. Han, "Scg: Saliency and contour guided salient instance segmentation," *IEEE TIP*, vol. 30, pp. 5862–5874, 2021.

[42] X. Tian, K. Xu, X. Yang, B. Yin, and R. W. Lau, "Learning to detect instance-level salient objects using complementary image labels," *IJCV*, vol. 130, no. 3, pp. 729–746, 2022.

[43] J. Pei, T. Cheng, H. Tang, and C. Chen, "Transformer-based efficient salient instance segmentation networks with orientative query," *IEEE TMM*, 2022.

[44] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*, 2014, pp. 345–360.

[45] J. Uhrig, M. Cordts, U. Franke, and T. Brox, "Pixel-level encoding and depth layering for instance-level semantic labeling," in *GCPR*, 2016, pp. 14–25.

[46] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, "Learning rgb-d feature embeddings for unseen object instance segmentation," in *Conference on Robot Learning*, 2021, pp. 461–470.

[47] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Processing: Image Communication*, vol. 38, pp. 115–126, 2015.

[48] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *IEEE CVPR*, 2012, pp. 454–461.

[49] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *ECCV*, 2014, pp. 92–109.

[50] D.-P. Fan, J. Zhang, G. Xu, M.-M. Cheng, and L. Shao, "Salient objects in clutter," *IEEE TPAMI*, vol. 45, no. 2, pp. 2344–2366, 2022.

[51] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE CVPR*, 2014, pp. 280–287.

[52] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[54] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *ICLR*, 2020.

[55] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," in *NeurIPS*, 2018.

[56] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[57] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *IEEE CVPR*, 2018, pp. 7794–7803.

[58] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection," *IEEE TIP*, vol. 31, pp. 6800–6815, 2022.

[59] N. Liu, N. Zhang, L. Shao, and J. Han, "Learning selective mutual attention and contrast for rgb-d saliency detection," *IEEE TPAMI*, vol. 44, no. 12, pp. 9026–9042, 2021.

[60] J. Pei, T. Cheng, D.-P. Fan, H. Tang, C. Chen, and L. Van Gool, "Osformer: One-stage camouflaged instance segmentation with transformers," in *ECCV*, 2022, pp. 19–37.

[61] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *IEEE CVPR*, 2019, pp. 6409–6418.

[62] Z. Cai and N. Vasconcelos, "Cascade r-cnn: high quality object detection and instance segmentation," *IEEE TPAMI*, vol. 43, no. 5, pp. 1483–1498, 2019.

[63] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *IEEE CVPR*, 2020, pp. 13 906–13 915.

[64] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *IEEE CVPR*, 2019, pp. 4974–4983.

[65] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *IEEE ICCV*, 2019, pp. 9157–9166.

[66] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," in *IEEE CVPR*, 2020, pp. 8573–8581.

[67] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *ECCV*, 2020, pp. 282–298.

[68] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *IEEE CVPR*, 2021, pp. 6910–6919.

[69] R. Guo, D. Niu, L. Qu, and Z. Li, "Sotr: Segmenting objects with transformers," in *IEEE ICCV*, 2021, pp. 7157–7166.

[70] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask transfiner for high-quality instance segmentation," in *IEEE CVPR*, 2022, pp. 4412–4421.

[71] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE ICCV*, 2017, pp. 2980–2988.

[72] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[73] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*, 2009, pp. 248–255.

[75] N. Huang, Q. Jiao, Q. Zhang, and J. Han, "Middle-level feature fusion for lightweight rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 6621–6634, 2022.

[76] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE ICCV*, 2021, pp. 10 012–10 022.

[77] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *CVMJ*, vol. 8, no. 3, pp. 415–424, 2022.

[78] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.